حاصلی، داود؛ فهیم‌نیا، فاطمه؛ نقشینه، نادر؛ عطاپور، هاشم؛ حسینی بهشتی، ملوک السادات (۱۳۹۸). مرور نظام‌مند پژوهش‌های حوزه گسترش پرس‌وجو در زبان فارسی. *پژوهشنامه کتابداری واطلاع‌رسانی*، ۹ (۱)، ۲۰۱– ۲۲۰.

مرور نظام‌مند پژوهش‌های حوزه گسترش پرس‌وجو در زبان فارسی

داود حاصلی[۱] فاطمه فهیم‌نیا[۲] نادرنقشینه[۳] هاشم عطاپور[۴] ملوک السادات حسینی بهشتی[۵]

تاریخ دریافت: ۱۳۹۷/۰۴/۱۶ تاریخ پذیرش: ۱۳۹۷/۰۵/۱۳ DOI: 10.22067/riis.v0i0.73983

چکیده

مقدمه: یکی از عوامل عدم موفقیت در بازیابی اطلاعات، ارائه نیاز اطلاعاتی کاربران در پرس‌وجوهای کوتاه و مبهم به نظام‌های اطلاعاتی است. گسترش پرس‌وجو با افزودن اصطلاحات مناسب به پرس‌وجوهای کاربران، راه‌حلی مناسب برای حل این مشکل است. هدف پژوهش حاضر مرور نظام‌مند متون پژوهشی گسترش پرس‌وجو در زبان فارسی است.

روش شناسی: فارسی و انگلیسی منابع اطلاعات علمی با کلیدواژه‌های مرتبط، تعداد ۳۵ اثر به زبان فارسی و ۱۸ اثر به زبان انگلیسی شناسایی شد. سپس با اعمال پالایش اولیه، معیارهای ورود و خروج از مطالعه و کنترل توسط متخصصان، تعداد شش اثر فارسی و هشت اثر انگلیسی برای ورود به مرور نظام‌مند انتخاب شدند. با طراحی کاربرگی، استخراج اطلاعات از آثار صورت پذیرفت. در ادامه، یافته‌های مرور نظام‌مند در پی دستیابی به چهار هدف پژوهش تحلیل شدند: شناسایی روش‌ها؛ شناسایی منابع دانشی؛ شناسایی مجموعه آزمون‌ها؛ و شناسایی شکاف‌های پژوهشی و ارائه پیشنهادهایی برای پژوهش‌های آینده در گسترش پرس‌وجوی زبان فارسی.

یافته‌ها: مرور پژوهش‌ها نشان داد ۱۴ اثر به گسترش پرس‌وجوی زبان فارسی پرداخته‌اند. این آثار براساس منابع دانشی اصطلاحات گسترش به چهار دسته تقسیم شدند: مبتنی بر ربط (هشت اثر)؛ مبتنی بر ساختارهای دانش (دو اثر)، مبتنی بر اطلاعات وب (دو اثر)، و مبتنی بر منابع ترکیبی (دو اثر). اغلب این پژوهش‌ها بر روی اسناد خبری انجام شده‌اند و از مجموعه آزمون روزنامه همشهری در نیمی از پژوهش‌ها به‌عنوان منبع دانشی اصطلاحات گسترش و نیز مجموعه آزمون استفاده شده است.

۱. دانشجوی دکتری علم اطلاعات و دانش‌شناسی دانشگاه تهران، dhaseli@ut.ac.ir

۲. دانشیار گروه علم اطلاعات و دانش‌شناسی دانشگاه تهران (نویسنده مسئول)، fahimnia@ut.ac.ir

۳. دانشیار گروه علم اطلاعات و دانش‌شناسی دانشگاه تهران، nnaghsh@ut.ac.ir

۴. استادیار گروه علم اطلاعات و دانش‌شناسی دانشگاه تبریز، hashematapour@tabrizu.ac.ir

۵. استادیار گروه علم اطلاعات و دانش‌شناسی پژوهشگاه علوم و فناوری اطلاعات (ایرانداک)، beheshti@irandoc.ac.ir

نتیجه‌گیری: تحقیقات حوزه گسترش پرس‌وجو در زبان فارسی نیازمند توسعه کمی با استفاده از روش‌های متنوع و به‌ویژه روش‌های مبتنی بر منابع ترکیبی است. منابع دانشی مختلف به‌ویژه هستی‌شناسی‌ها و منابع وب می‌بایست برای گسترش پرس‌وجو در زبان فارسی مورد توجه و استفاده قرار گیرند. همچنین استفاده از مجموعه آزمون‌های استاندارد برای پژوهشگران این امکان فراهم می‌کند که بتوانند روش‌های مختلف را با هم مقایسه کنند.

کلیدواژه‌ها: گسترش پرس‌وجو، زبان فارسی، بازخورد ربط، ساختارهای دانش، اطلاعات وب

مقدمه

روزانه طیف وسیعی از افراد با نیازها و انگیزه‌های مختلف به جستجوی اطلاعات در نظام‌های بازیابی اطلاعات می‌پردازند. کاربران به‌منظور برقراری ارتباط با نظام‌های بازیابی اطلاعات، نیازهای اطلاعاتی‌شان را در قالب پرس‌وجوها فرمول‌بندی می‌نمایند (Phan, Bailey & Wilkinson, 2007). معمولاً خلاصه‌ای از نیاز اطلاعاتی کاربر توسط پرس‌وجویی مشتمل بر مجموعه‌ای از کلمات کلیدی بیان می‌شود (شبان‌زاده حبیب‌آبادی، ۱۳۸۹). اغلب پرس‌وجوهای ارائه‌شده توسط کاربران کوتاه و مبهم هستند. نخستین مشکل پرس‌وجوهای کاربران در وب تعداد کم اصطلاحات پرس‌وجو است. پژوهش‌ها نشان می‌دهند کاربران تمایل به ارائه پرس‌وجوهای کوتاه دارند زیرا برای مشخص ساختن نیازهای اطلاعاتی خود با کمبود دانش موضوعی مواجه هستند، و طول یک پرس‌وجوی تحت وب بین ۲ تا ۳ اصطلاح است (Spink, Wolfram, Jansen, & Saracevic, 2001; Wollersheim, 2005). تعداد کم اصطلاحات پرس‌وجو باعث می شود اصطلاحات مهمی که توصیفگر نیاز اطلاعاتی هستند، در پرس‌وجو ظاهر نشوند (Spink, et al., 2001) و معنی مناسب و کافی برای پرس‌وجوی مورد نظر فراهم نیاید. مشکل دیگر پرس‌وجوها ابهام است. برخی از عوامل ایجادکننده ابهام در پرس‌وجوهای کاربران برای نظام‌های بازیابی عبارتند از: فرمول‌بندی ضعیف پرس‌وجو، اصطلاحات مترادف و اصطلاحات دارای تعدد معانی، و عملکرد نادرست نظام (Zhang, 2013). برای کاربران معمولی که قادر به بیان نیازهای اطلاعاتی در قالب یک پرس‌وجوی مؤثر نیستند، فرمول‌بندی ضعیف پرس‌وجو یک دلیل ابتدائی برای پرس‌وجوهای مبهم محسوب می‌شود (Harman & Buckley, 2004). وجود اصطلاحات مترادف برای یک مفهوم باعث می شود همه اسناد مرتبط با آن مفهوم بازیابی نشوند و بازیافت جستجو کاهش یابد. همچنین مشکل وجود اصطلاحات دارای تعدد معانی در پرس‌وجو، احتمال بازیابی اسناد نامرتبط با معناهای دیگر که مورد نظر کاربر نیستند را افزایش می‌دهد که این امر دقت نتایج بازیابی را کاهش می‌دهد (Zhang, 2013). موضوع رایج دیگر در فرمول‌بندی پرس‌وجو استفاده از اصطلاحات اشتباه و غلط است، ممکن است کاربران

اصطلاحات را از نظر املائی اشتباه بنویسند یا به‌دلیل فقدان دانش موضوعی کافی اصطلاحات نامناسب به‌کار ببرند (شبان‌زاده حبیب‌آبادی، ۱۳۸۹).

یکی از رویکردهای رایج برای حل مشکل پرس‌وجوهای کوتاه و مبهم که از دهه ۱۹۶۰ میلادی تاکنون مورد استفاده قرار می‌گیرد گسترش پرس‌وجو[1] است. گسترش پرس‌وجو رویکردی پذیرفته شده است که به‌صورت گسترده مورد استفاده قرار می‌گیرد و پرس‌وجوهای کوتاه کاربران را با افزودن اصطلاحات اضافی از بافت، تقویت می‌کند، همچنین با محدود ساختن معنی واژگان به‌وسیله اصطلاحات اضافه‌شده به پرس‌وجو، مشکل ابهام در زبان طبیعی را نیز حل می‌کند (Zhang, 2013). از جمله عوامل موفقیت استفاده از گسترش پرس‌وجو در بازیابی اطلاعات، حل مشکل عدم تطابق واژگان موجود در پرس‌وجوی کاربر با واژگان موجود در مجموعه اسناد است (Wollersheim, 2005).

برخی گسترش پرس‌وجو را اصلاح مجدد پرس‌وجوی کاربر با افزودن اصطلاحات اضافی و وزن‌دهی مجدد اصطلاحات پرس‌وجو توسط نظام می‌دانند (Lavrenko & Croft, 2017; Lee, Croft, & Allan, 2008) برخی نیز تنها بر وزن‌دهی مجدد اصطلاحات پرس‌وجو تمرکز می‌کنند (Bendersky & Croft, 2008; Robertson & Jones, 1976). برخی نیز سه رویکرد در نظر می‌گیرند، افزودن اصطلاحات؛ وزن‌دهی مجدد؛ و ترکیبی از افزودن اصطلاحات اضافی و وزن‌دهی مجدد به آنها (Baeza-Yates, & Ribeiro-Neto, 1999). روش‌ها، فنون و الگوریتم‌های مختلفی برای گسترش پرس‌وجوهای کاربران در نظام‌های بازیابی اطلاعات به‌کار رفته‌اند. فرایند گسترش پرس‌وجو به سه گروه دستی، خودکار، و تعاملی تقسیم می‌شود (Abdelmgeid Amin, 2008; Zhang, 2013). در گسترش پرس‌وجوی دستی[2] کاربر بر پایه تجربه و دانش خود از حوزه موضوعی و مجموعه اسناد، اصطلاحات گسترش را تعیین می‌کند (Bhogal, MacFarlane, & Smith, 2007). گسترش پرس‌وجوی دستی برای متخصصان موضوعی و حرفه‌ای مناسب است. در فرایند گسترش پرس‌وجوی تعاملی،[3] نظام مجموعه‌ای از اصطلاحات بالقوه گسترش پرس‌وجو را شناسایی و برای کاربر ارائه می‌کند و کاربر تصمیم می‌گیرد چه اصطلاح یا اصطلاحاتی برای گسترش مناسب است. پیش‌فرض گسترش پرس‌وجوی تعاملی این است که افراد نسبت به ماشین بیشتر قادر به قضاوت ربط و مفید بودن اصطلاحات هستند و این روش اثربخشی را در کنش و

1. Query Expansion
2. Manual Query Expansion
3. Interactive Query Expansion

عمل نشان می‌دهد (Azad & Deepak, 2019). در گسترش پرس‌وجوی خودکار[1] نظام به‌صورت خودکار و بدون هیچ مداخله‌ای از کاربر اصطلاحات گسترش را انتخاب می‌کند و پرس‌وجو را برای کاربر فرمول‌بندی مجدد می‌نماید. گسترش اصطلاحات می‌تواند نتیجه منابع متنوعی باشد، شامل مجموعه‌های متنی، فرهنگ‌ها، و پایه‌های دانش.

روش‌ها و الگوریتم‌های گسترش پرس‌وجو در زبان‌های مختلف پیاده سازی و مورد استفاده قرار گرفته‌اند. بررسی متون منتشرشده نشان می‌دهد چندین روش و الگوریتم گسترش پرس‌وجو در زبان فارسی انجام شده است. زبان فارسی که از شاخه زبان‌های هند و اروپایی به‌شمار می‌رود یکی از زبان‌های مهم در آسیا است. در کشورهای ایران و تاجیکستان، فارسی زبان رسمی بوده و در کشور افغانستان در کنار زبان پشتو، یکی از دو زبان رسمی است. همچنین زبان فارسی، زبان رسمی کشور هندوستان تا پیش از ورود استعمار انگلیس بوده است. امروزه حجم زیادی از تولیدات علمی و صفحات وب به زبان فارسی تولید می‌شود. نمایه نزدیک به یک میلیون مقاله مجلات فارسی در پایگاه مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری،[2] نمایه حدود ۷۵۰,۰۰۰ مقاله کنفرانسی در پایگاه سیویلیکا،[3] و نمایه ۶۰۰,۰۰۰ پایان‌نامه در پایگاه گنج ایرانداک[4] نمونه‌هایی از مدارک علمی تولید شده به زبان فارسی در محیط وب هستند. همچنین برای سازماندهی و بازیابی این اطلاعات در وب فارسی، پژوهش‌های زیادی در حوزه بازیابی اطلاعات و پردازش زبان طبیعی در زبان فارسی صورت گرفته و یا در حال انجام است.

بررسی متون منتشرشده نشان می‌دهد با توجه به حجم تحقیقات انجام شده در خصوص گسترش پرس‌وجو در زبان‌های مختلف (مانند زبان انگلیسی، چینی و فرانسه)، پژوهش‌های اندکی بر روی گسترش زبان فارسی انجام‌شده است. علاوه بر این، تصویری جامع از اندک پژوهش‌های انجام شده، روش‌های به‌کار رفته و منابع مورد استفاده در آنها وجود ندارد. بر این اساس، پژوهش حاضر به مرور نظام‌مند تحقیقات گسترش پرس‌وجو در زبان فارسی می‌پردازد. در این مقاله، با طبقه‌بندی پژوهش‌های صورت گرفته در گسترش پرس‌وجو از نظر منابع دانشی اصطلاحات گسترش، از ارائه مشخصات صرفاً فنی اجتناب شده است؛ زیرا بهبود روش‌ها و الگوریتم‌های گسترش پرس‌وجو بدون در نظر گرفتن منابع گسترش پرس‌وجو منجر به پیشرفت قابل توجهی نخواهد شد. این مرور نظام‌مند می‌تواند انباشت علمی مطالعات

1. Automatic Query Expansion
2. http://search.ricest.ac.ir/Inventory/index_10702.htm
3. https://www.civilica.com/index.php
4. https://ganj.irandoc.ac.ir

حوزه گسترش پرس‌وجو در زبان فارسی را از لحاظ استفاده از منابع، روش‌ها، فنون، مجموعه آزمون و ... به تصویر بکشد و با ترکیب مطالعات پیشین و بررسی کارهای انجام‌شده، خلأهای پژوهشی در این حوزه را به پژوهشگران عرضه نماید. این پژوهش می‌تواند نقطه عزیمت مهمی برای پژوهشگران حوزه علم اطلاعات و دانش‌شناسی به‌منظور توسعه شناختی روش‌های گسترش پرس‌وجو و ایجاد منابع دانشی مختلف برای تأمین اصطلاحات گسترش پرس‌وجو و نیز ساخت مجموعه آزمون‌هایی برای ارزیابی نظام‌های بازیابی اطلاعات در زبان فارسی باشد.

پژوهش مروری نظام‌مند حاضر در پی تحقق چهار هدف ذیل در حوزه تحقیقات گسترش پرس‌وجوی زبان فارسی است: ۱) شناسایی روش‌ها و الگوریتم‌های استفاده‌شده؛ ۲) شناسایی منابع دانشی اصطلاحات گسترش؛ ۳) شناسایی مجموعه آزمون‌های استفاده شده برای گسترش پرس‌وجو؛ و ۴) شناسایی شکاف‌های پژوهشی و ارائه پیشنهادهایی برای پژوهش‌های آینده.

روش‌شناسی

پژوهش حاضر با استفاده از روش مرور نظام‌مند انجام‌شده است. در مرور نظام‌مند با شناسایی دقیق، منظم و برنامه‌ریزی‌شده تمام مطالعات مرتبط، می‌توان نقد عینی‌تری انجام داد و به مشکلات مربوط به مرورهای دیگر مانند مرور نقلی فائق آمد (ملبوس‌باف و عزیزی، ۱۳۸۹). در این پژوهش تمامی مقالات منتشر شده مرتبط با گسترش پرس‌وجوی زبان فارسی مورد بررسی قرار گرفته است. در پژوهش حاضر جهت اطمینان از کامل بودن و ثبات در مرور نظام‌مند پژوهش‌ها، از راهنمای ارائه شده توسط اُکلی و شابرم (Okoli & Schabram, 2010) استفاده شده است. این راهنما برای اطمینان از دقت زیاد در انجام روش‌شناسی در زمان انجام یک مرور نظام‌مند طراحی شده است. براساس این راهنما مرور نظام‌مند پژوهش حاضر در شش گام (جدول ۱) طراحی شد. در ادامه شش گامِ اجراییِ مرور نظام‌مند توضیح داده شده است.

۱. شناسایی نیاز به مرور پژوهش‌ها: بررسی متون نشان داد هیچ پژوهشی در خصوص مرور پژوهش‌ها در زمینه گسترش پرس‌وجو در زبان فارسی انجام نشده است. این در حالی است که در زبان انگلیسی و زبان‌های دیگر، پژوهش‌های مختلفی با استفاده از روش مرور نظام‌مند، نه‌تنها برای حوزه کلی گسترش پرس‌وجو بلکه برای زیر حوزه‌های جزئی آن صورت گرفته است.

۲. تدوین اهداف مرور نظام‌مند: با انتشار پژوهش‌های گسترش پرس‌وجوی زبان فارسی در چند سال اخیر، لازم است روش‌ها، منابع دانشی و مجموعه آزمون‌های استفاده‌شده برای آن شناسایی شود و خلأهای پژوهشی آن مشخص گردد.

۳. جستجو در منابع: جستجوی اینترنتی در پایگاه‌های داخلی شامل مقالات کنفرانس‌ها و همایش‌ها[1]، جویشگر علم‌نت[2]، بانک نشریات کشور[3]، پایگاه اطلاعات علمی جهاد دانشگاهی[4] و پایگاه پژوهشگاه علوم و فناوری اطلاعات[5] به منظور یافتن منابع فارسی صورت گرفت. همچنین بررسی منابع و مآخذ آثار نیز در دستیابی به برخی از منابع راهگشا بود. برای جستجو در منابع فارسی از کلیدواژه‌های «گسترش»، «بسط»، «پرس‌وجو»، «جستجو»، و «پرسش» با ترکیب‌های مناسب و با نگارش‌های املائی متفاوت استفاده شد. در مورد مقالات انگلیسی نیز پایگاه‌های خارجی ساینس‌دایرکت[6] و موتور کاوش علمی گوگل اسکالر[7] با ردگیری آثار استنادکننده و نیز منابع و مآخذ آثار مورد بررسی قرار گرفتند. کلیدواژه‌های استفاده شده برای جستجو در زبان انگلیسی شامل «Query Expansion»، «Persian»، و «Farsi» بودند. در این گام ۳۵ اثر از پایگاه‌های فارسی و ۱۸ اثر به از پایگاه‌های خارجی به زبان انگلیسی بازیابی شد.

٤. گزینش منابع مرور: در پالایش اولیه ۲۱ اثر به زبان فارسی و ۱۸ اثر به زبان انگلیسی گزینش شدند. محدودیت زمانی برای پژوهش در نظر گرفته نشده است. معیار ورود به مطالعه برای آثار، عبارتند بودند از: ۱) گسترش پرس‌وجو در زبان فارسی و ۲) اعمال الگوریتم‌های گسترش پرس‌وجو و بهبود بازیابی. معیار خروج از مطالعه نیز شامل حذف قالب‌های تکراری یک اثر (مانند انتشار یک عنوان در قالب پایان‌نامه، مقاله مجله یا مقاله کنفرانس) بود. از ۲۱ اثر به زبان فارسی، تنها هفت اثر در خصوص اعمال الگوریتم‌های گسترش پرس‌وجو در زبان فارسی بود، ۱۲ پژوهش به گسترش پرس‌وجوی زبان انگلیسی پرداخته بودند و دو اثر با استفاده از روش پیمایش و به‌صورت نظرسنجی انجام شده بود که از مطالعه کنار گذاشته شدند. از ۱۸ اثر انگلیسی، هشت اثر مرتبط با گسترش پرس‌وجو در زبان فارسی تشخیص داده شدند. در مجموع هفت اثر به زبان فارسی و هشت اثر به زبان انگلیسی برای بررسی مرتبط شناخته شدند. آثار انتخاب شده

1. www.civilica.com
2. www.elmnet.ir
3. www.magiran.com
4. www.sid.ir
5. www.ganj.irandoc.ac.ir
6. www.ScienceDirect.com
7. www.scholar.google.com

توسط یک نفر متخصص و صاحب‌نظر کنترل و بررسی شد. یک اثر به‌دلیل تشابه کامل با یک اثر دیگر توسط متخصص کنار گذاشته شد. در مجموع ۱۴ اثر برای مرور نظامند انتخاب شدند.

۵. استخراج اطلاعات آثار: کاربرگی جهت استخراج اطلاعات بر اساس اهداف پژوهش از آثار، طراحی و تهیه شد که در آن تعیین شده بود چه اطلاعاتی از کل اثر باید استخراج شود.

٦. تجزیه و تحلیل و ارائه یافته‌ها: در گام نهایی نیز تجزیه و تحلیل و ارائه یافته‌های پژوهش انجام شد که حاصل کل مرور نظام‌مند است. جدول ۱ گام‌های مرور نظام‌مند پژوهش حاضر را نشان می‌دهد.

جدول ۱. مراحل مرور نظام‌مند

گام	فرایند مرور نظام‌مند	تعداد مقالات باقیمانده	
		فارسی	انگلیسی
گام ۱	شناسایی نیاز به مرور پژوهش‌ها	–	–
گام ۲	تدوین اهداف مرور نظام‌مند	–	–
گام ۳	جستجو در منابع [الکترونیکی فارسی و انگلیسی]	۳۵	۱۸
گام ۴	پالایش اولیه آثار	۲۱	۱۲
	اعمال معیارهای ورود و خروج از مرور نظام‌مند	۷	۸
	کنترل و بررسی آثار انتخاب‌شده توسط متخصص و صاحب‌نظر	۶	۸
گام ۵	استخراج اطلاعات آثار	۶	۸
گام ۶	تجزیه و تحلیل و ارائه یافته‌ها	۶	۸

یافته‌ها

زبان و قالب انتشار تحقیقات: در نهایت شش اثر به زبان فارسی و هشت اثر به زبان انگلیسی وارد مرور نظام‌مند شدند. از این تعداد قالب شش اثر مقاله نشریه، پنج اثر مقاله کنفرانسی و سه اثر پایان‌نامه کارشناسی ارشد بودند (جدول ۲).

جدول ۲. زبان و قالب آثار مورد بررسی

		قالب آثار			کل
		مقاله نشریه	مقاله کنفرانسی	پایان‌نامه	
زبان آثار	فارسی	۲ (۱۴٪/۳)	۱ (۷٪/۱)	۳ (۲۱٪/۴)	۶ (۴۲٪/۹)
	انگلیسی	۴ (۲۸٪/۶)	۴ (۲۸٪/۶)	۰	۸ (۵۷٪/۱)
کل		۶ (۴۲٪/۹)	۵ (۳۵٪/۷)	۳ (۲۱٪/۴)	۱۴

روش‌ها و رویکردهای گسترش پرس‌وجو در زبان فارسی: در مرور نظام‌مند حاضر روش‌های گسترش پرس‌وجو در زبان فارسی براساس نوع منبع اصطلاحات گسترش به چهار نوع دسته‌بندی شده‌اند: مبتنی بر ربط (هشت اثر)، مبتنی بر ساختارهای دانش (دو اثر)، مبتنی بر اطلاعات وب (دو اثر)، و مبتنی بر منابع ترکیبی (دو اثر).

گسترش پرس‌وجو مبتنی بر ربط: این نوع گسترش پرس‌وجو به دو نوعِ بازخورد ربط و بازخورد شبه ربط[1] صورت می‌گیرد. بازخورد ربط، فرایند جستجو را به‌عنوان یک عملیات تعاملی در نظر می‌گیرد و اسنادی را که کاربران مرتبط تشخیص می‌دهند برای گسترش پرس‌وجوها انتخاب می‌کند (Azad & Deepak, 2019). در رویکرد بازخورد شبه ربط، اسناد با رتبه بالا در فهرست نتایج پرس‌وجوی اولیه به‌عنوان اسناد مرتبط شناخته می‌شوند. زمانی‌که بازخورد ربط کاربر در دسترس نباشد این رویکرد که به‌عنوان بازخورد کور نیز شناخته می‌شود، مورد استفاده قرار می‌گیرد. بسیاری از نظام‌های بازیابی اطلاعات از این رویکرد برای گسترش پرس‌وجو بهره می‌برند (Atwan & Mohd, 2017). سه الگوریتم گسترش پرس‌وجو مبتنی بر ربط وجود دارد: روکیو[2]، احتمالاتی[3]، و تحلیل بافت محلی[4]. الگوریتم روکیو براساس مدل فضای برداری است (Salton, Wong, & Yang, 1975) و هدف آن انتقال بردار پرس‌وجوی گسترش‌یافته به نزدیک میانگین بردار اسناد مرتبط و دور کردن از میانگین اسناد غیرمرتبط است (Rocchio, 1971). الگوریتم احتمالاتی، با استفاده از وزن‌های اصطلاح براساس رخدادهای اصطلاح در اسناد مرتبط و اسناد نامرتبط محاسبه می‌شود و یک سند را به‌وسیله احتمال ربط سند با پرس‌وجو رتبه‌بندی می‌کند (Robertson & Jones, 1976). در تحلیل بافت محلی فرض این است که یک سند طولانی ممکن است چندین موضوع را پوشش دهد و تنها یک موضوع آن سند مورد نظر کاربر باشد. بنابراین محتواهایی

1. Pseudo Relevance Feedback
2. Rocchio
3. Probabilistic
4. Local Context Analysis (LCA)

با هم‌رخدادی متوالی در فاصـــله نزدیک با اصـــطلاحات پرس‌وجو، منابع قابل اعتماد برای اصـــطلاحات پرس‌وجو هستند زیرا اصطلاحات موجود در بافت‌های مشابه، اغلب دارای معانی مشابه هستند (Agichtein & Cucerzan, 2005). در اسناد فارسی یک اثر با رویکرد بازخورد و هفت اثر با رویکرد بازخورد شبه ربط انجام شده است.

بازخورد ربط: در زبان فارسـی یک اثر با اسـتفاده از رویکرد بازخورد ربط و اسـتفاده از مجموعه آزمون حوزه حقوق انجام شده است. صبوری، بشیری و اروم‌چیان (Saboori, Bashiri & Oroumchian, 2008) با استفاده از بازخورد ربط به تعیین تأثیر مدل روکیو در وزن‌دهی مجدد پرس‌وجو برای بازیابی اسناد فارسی پرداختند. آزمایش آنها در مجموعه آزمون قوانین با موفقیت همراه بود.

بازخورد شبه ربط: تاکنون بیشترین آثار (هفت اثر) با استفاده از این رویکرد به گسترش پرس‌وجو در زبان فارسی پرداخته‌اند. در این هفت اثر از مجموعه آزمون همشهری به‌عنوان منبع دانشی اصطلاحات گسترش و نیز مجموعه آزمون استفاده شده است. مجموعه همشهری پیکره‌ای است حاوی ۳۱۸ هزار سند مربوط به اخبار سـال‌های ۱۳۷۵ تا ۱۳۸۶ که با خزش وب‌سـایت همشـهری و چندین مرحله پیش‌پردازش و برچسب‌گذاری حاصل آمده است. همه اسناد مجموعه همشهری دارای برچسب هستند که نشان می‌دهد هر سـند در چه رده‌ای (اقتصـادی، سـیاسـی و...) اسـت (AleAhmad, Amiri, Darrudi, Rahgozar, & Oroumchian, 2009).

دولامیک و سـاوُی (Dolamic & Savoy, 2009) از گسـترش پرس‌وجو مبتنی بر روش بازخورد شـبه ربط و مدل روکیو برای گسـترش پرس‌وجو اسـتفاده نمودند. آزمایش آنها بر روی مجموعه آزمون همشهری و بخش فارسی مجموعه اسناد کِلِف[1] با مدل‌های مختلف بازیابی اطلاعات (DFR، اُکاپی و مدل زبانی)، استراتژی‌های نمایه سازی (اصطلاح با ریشه‌یابی و بدون ریشه‌یابی و ۵ گِرم)، راهبردهای گسترش پرس‌وجو (روکیو، بر پایه معکوس فراوانی اسناد یا بدون گسترش) و شکل‌دهی پرس‌وجو (پرس‌وجوهای کوتاه، متوسـط و طولانی) انجام شـد. آنها نشـان دادند شـکل‌دهی پرس‌وجو با حذف تمام پسـوندها و پیشوندهای واژه‌های زبان فارسی و نیز پرس‌وجوهای طولانی که از ۱۰ سند مرتبط ۲۰ اصطلاح گسترش به پرس‌وجو اضافه می‌کند، بازیابی بهبود می‌دهد.

دو پژوهش با اسـتفاده از رویکرد بازخورد شـبه ربط و مدل احتمالاتی در زبان فارسـی انجام شـده است. کریسانی، رهگذر و اروم‌چیان (Karisani, Rahgozar & Oroumchian, 2016) یک روش ساده و

1. CLEF

در عین حال کاربردی از روش بازخورد شبه ربط به‌منظور شناسایی اصطلاحات مهم و دارای بار اطلاعاتی بیشتر در پرس‌وجو و وزن‌دهی مجدد به آنها با استفاده از شباهت اسناد (الگوریتم احتمالاتی) ارائه دادند. یافته‌ها در مجموعه‌های آزمون استاندارد فارسی همشهری ۱ و همشهری ۲ و بخش فارسی مجموعه آزمون انگلیسی فایر[1] موفقیت این رویکرد را نشان داد. هاشمی و شاکری (Hashemi & Shakery, 2014) به‌منظور ساخت مدل گسترش پرس‌وجو، برای هر پرس‌وجو ۱۰ سند بازیابی شده با رتبه بالا در مجموعه آزمون همشهری و اخبار بی‌بی‌سی را برای افزودن ۱۰۰ اصطلاح مورد استفاده قرار دادند. یافته‌ها نشان داد گسترش پرس‌وجو بر اساس اصطلاحات مرتبط مؤثر بوده است.

چهار پژوهش با استفاده از رویکرد بازخورد شبه ربط و تحلیل بافت محلی انجام شده است. آل‌احمد، حکیمیان، مهدی‌خانی و ارومچیان (Aleahmad, Hakimian, Mahdikhani & Oroumchian, 2007) به ارزیابی مدل فضای برداری مبتنی بر اصطلاح و ان‌گرم و روش گسترش پرس‌وجوی بازخورد شبه ربط با استفاده از طرح‌های وزن‌دهی مختلف پرداختند. آنها تعداد ۱۰ اصطلاح نخست رتبه‌بندی شده از ۲۰ نتیجه اول بازیابی‌شده از مجموعه آزمون همشهری را به هر پرس‌وجو اضافه نمودند. یافته‌های آزمون نشان داد گسترش پرس‌وجو با استفاده از روش تحلیل بافت محلی مؤثر بوده است و گسترش پرس‌وجوی زبان فارسی با ۴ گرم نتایج بهتری در بردارد. حکیمیان و تقی‌یاره (Hakimian & Taghiyareh, 2007) مجموعه همشهری را برای تطبیق سه پارامتر ۱) تعداد مفاهیم برای گسترش پرس‌وجو (۱۰ الی ۳۰ مفهوم)؛ ۲) تعداد اسناد بازیابی‌شده اولیه برای بازخورد محلی (۳۰ الی ۱۰۰ سند)؛ و ۳) تعداد اصطلاحات برای کشف مفاهیم و وزن‌دهی، برای گسترش پرس‌وجو مورد استفاده قرار دادند. آنها نشان دادند زمانی‌که ۲۰ مفهوم برای گسترش پرس‌وجو مورد استفاده قرار می‌گیرد نقطه بهینه محسوب می‌شود، با این حال، افزایش دو پارامتر دیگر نیز در اکثر موارد نتایج را بهبود می‌دهد. حکیمیان و تقی‌یاره (Hakimian & Taghiyareh, 2008) در راستای پژوهش پیشین، تعداد مفاهیم را تا ۶۰ مفهوم افزایش دادند. یافته‌های آنها نشان داد افزایش مفاهیم گسترش به ۲۵، ۳۰، ۳۵، ۵۰ مفهوم باعث بهبود عملکرد بازیابی نمی‌شود اما افزایش مفاهیم به ۴۰ و ۴۵ مفهوم باعث بهبود عملکرد بازیابی مؤثر می‌شود. خالقی و مینایی (۱۳۹۴) با استفاده از بررسی هم‌رخدادی اصطلاحات در پارگراف‌ها، اقدام به ساخت مجموعه‌های هم‌رخدادی برای هر اصطلاح نموده و از آن‌ها برای گسترش پرس‌وجو با رویکرد تحلیل بافت محلی استفاده کردند.

1 Forum for Information Retrieval Evaluation

یافته‌های آزمایش‌ها در مجموعه آزمون هم‌شهری (۱۳۷۵-۱۳۸۶) نشان داد چارچوب پیشنهاد شده بازیابی را بهبود می‌دهد.

گسترش پرس‌وجو مبتنی بر ساختارهای دانش: در این روش، اصطلاحات گسترش از ساختارهای دانش و با استفاده از دو رویکرد وابسته به پیکره و مستقل از پیکره استخراج می‌شود. رویکرد وابسته به پیکره، اصطلاحات را از مجموعه‌های متنوعی مانند خوشه‌بندی اصطلاحات و اصلاحنامه‌های خودکار مستخرج از متن تأمین می‌کند (Atwan & Mohd, 2017). رویکرد مستقل از پیکره از منابع خارجی همچون واژه‌نامه‌ها، اصطلاحنامه‌های عمومی، اصطلاحنامه‌های حوزه‌های خاص و هستی‌شناسی‌ها استفاده می‌کند (Efthimiadis, 1996). وردنت عمومی زبان فارسی که فارس‌نت[1] نام دارد، نخستین شبکه واژگان یا هستی‌شناسی زبان فارسی و پایگاه دانشی است که حاوی اطلاعات در مورد واژه‌ها و ترکیبات زبان (مفاهیم)، اطلاعات نحوی آنها و روابط معنایی میان آنهاست (Shamsfard, et al., 2010).

گسترش پرس‌وجو در زبان فارسی با استفاده از ساختارهای وابسته به پیکره انجام نشده است. دو پژوهش با وردنت فارسی و واژه‌نامه صورت گرفته است که جزو ساختارهای مستقل از پیکره هستند. ساعدی (۱۳۹۰) به گسترش پرس‌وجو با استفاده از روابط موجود در بین مفاهیم هستی‌شناسی عمومی فارسی‌نت اقدام نمود. وی هر پرس‌وجو را با یک تا چهار اصطلاح از هستی‌شناسی فارسی‌نت گسترش داده است. بهترین عملکرد و افزایش دقت با افزودن سه اصطلاح به پرس‌وجوی اولیه به‌دست آمده است. دیانت، علی‌احمدی، اخلاقی، باباعلی (۱۳۹۵) از گسترش پرس‌وجو به‌عنوان پیش‌پردازشی برای بهبود بازیابی اطلاعات حاصل از بازشناسی گفتار استفاده نمودند. آنها با استفاده از یک روش بازیابی برداری در مجموعه آزمون دادگان فارس‌دات بزرگ، موفقیت گسترش پرس‌وجو را در بهبود بازشناسی گفتار نشان دادند.

گسترش پرس‌وجو مبتنی بر اطلاعات وب: این رویکرد از اسناد وب، پایگاه‌های آنلاین دانش (مانند ویکی‌پدیا) یا لاگ پرس‌وجوها به‌عنوان منبع اصطلاحات گسترش استفاده می‌کند. انگیزه برای استفاده از اطلاعات وب برای گسترش پرس‌وجو، غنی‌ساختن مجموعه با استفاده از اطلاعات خارجی است، اطلاعات خارجی پویا و نشان‌دهنده دیدگاه عموم هستند (Zhang, 2013). ویکی‌پدیا بزرگ‌ترین دانشنامه وب چندزبانه است که به‌عنوان یک منبع اطلاعات ساختاریافته منشأ تولید بسیاری از ابزارهای بازیابی اطلاعات

1. Farsnet

و پردازش زبان طبیعی مانند هستی‌شناسی‌های دی‌بی‌پدیا[1]، یاگو[2] و WEX است (Mehdi, Okoli, Mesgari, Nielsen, & Lanamäki, 2017). همچنین لاگ‌های پرس‌وجوی تولیدشده توسط موتورهای کاوش وب، تعاملات کاربران را شامل تجربه‌های آنها در فرمول‌بندی مجدد پرس‌وجوها و میزان دستیابی به نتایج مطلوب در قالب نشست‌های پرس‌وجو ثبت می‌کنند (Azad & Deepak, 2019). در زبان فارسی، در دو اثر روابط معنایی ویکی‌پدیا و لاگ‌های پرس‌وجوی موتور کاوش گوگل برای گسترش پرس‌وجو مورد استفاده قرار گرفته‌اند.

پایگاه‌های دانش آنلاین: فرهودی، محمودی، زارع‌بیدکی، یاری و آزادنیا (Farhoodi, Mahmoudi, Zare Bidoki, Yari & Azadnia, 2009) با استفاده از روابط معنایی مفاهیم موجود در ویکی‌پدیا و ساختار اسناد موجود در آنکه براساس گراف موضوعی است یک هستی‌شناسی فارسی ایجاد کردند. آنها با استفاده از هستی شناسی و روابط میان مفاهیم، اصطلاحات گسترش را وزن‌دهی کردند و به موتور کاوش گوگل ارسال کردند. سپس ۲۰ نتیجه نخست را با کمک کاربران متخصص حوزه رایانه مورد قضاوت ربط قرار دادند. یافته ها حاکی از افزایش میزان دقت نتایج بازیابی‌شده پرس‌وجوی گسترش یافته با کمک هستی‌شناسی بود.

لاگ پرس‌وجوها: خسروی، فتاحی، پریرخ و دیانی (۱۳۹۲) با استفاده از پرس‌وجوهای ثبت شده پیشین کاربران (لاگ پرس‌وجو) که در کلیدواژه‌ها و پیشنهادات موتور کاوش گوگل نمایان می‌شوند، اقدام به گسترش پرس‌وجو نمودند. یافته‌های پژوهش آنها، افزایش میزان ربط نتایج بازیافتی را گزارش می‌کند. در این رویکرد علاوه بر اصطلاحات از عبارت‌ها نیز برای گسترش پرس‌وجو استفاده شده است.

گسترش پرس‌وجو با استفاده از منابع ترکیبی: روش‌های ترکیبی در گسترش پرس‌وجو، دو یا چند روش را برای ایجاد روشی مؤثرتر ادغام می‌کنند تا از این طریق بر نقاط ضعف روش‌های گسترش پرس‌وجوی خودکار غلبه نمایند. ویژگی‌های پرس‌وجو شامل اندازه پرس‌وجو، طول اصطلاحات، مسائل لغوی، ابهام، دشواری و هدف آن باعث می‌شود نیاز به روش خاصی برای گسترش هر یک از این ویژگی‌ها باشد (Atwan & Mohd, 2017). کریسانی (۱۳۹۰) تشکیل اصطلاحات اولیه گسترش از نقش و قابلیت شبکه معنایی (روش مبتنی بر ساختار دانش- مستقل از پیکره) و برای وزن‌دهی به آنها از مجموعه اسنادی که حاصل از بازخورد ربط (روش مبتنی بر ربط) هستند، استفاده نموده است. گسترش پرس‌وجو از طریق

1. DBpedia
2. YAGO

ترجمه کلمات به انگلیسی و گسترش آنها با استفاده از وردنت[1] انگلیسی و ترجمه دوباره کلمات گسترش‌یافته به فارسی انجام شده است. عبدالحسینی (۱۳۹۲) با اعمال پرس‌وجوی اولیه و بازیابی اسناد با رتبه بالا، یک گراف ارتباط مفهومی با استفاده از روش‌های آماری مفاهیم اصلی اسناد و ارتباط میان آن‌ها ایجاد نمود. با استفاده از گراف مذکور (روش مبتنی بر اطلاعات وب- اسناد وب) و هستی‌شناسی فارس‌نت (روش مبتنی بر ساختار دانش- مستقل از پیکره)، گروه‌بندی معنایی اصطلاحات صورت گرفت و اصطلاحات گسترش استخراج شدند. وی برای انتخاب اصطلاحات پرس‌وجو، در گراف با استفاده از محاسبات آماری وزنی و در هستی‌شناسی از الگوریتم ژنتیک با روش ترکیبی هم‌رخدادی و روش‌های مبتنی بر بسامد واژه استفاده نموده است. در این پژوهش پرس‌وجوهای مجموعه دادگان همشهری مورد استفاده قرار گرفته و یافته‌ها نشان می‌دهد میانگین متوسط دقت برای روش گراف نتایج بهتری نسبت به هستی‌شناسی دارد.

جدول ۳ منابع دانشی اصطلاحات پرس‌وجو، مجموعه آزمون و نوع قضاوت ربط را نشان می‌دهد. حدود نیمی از پژوهش‌های گسترش پرس‌وجو در زبان فارسی از مجموعه روزنامه همشهری به‌عنوان منبع دانشی اصطلاحات گسترش و همچنین مجموعه آزمون استفاده نمودند. همچنین برخی از مجموعه آزمون‌های خارجی همچون کلف، فایر و اخبار بی‌بی‌سی دارای اسناد و پرس‌وجوهای فارسی استاندارد هستند که برای وظایف بازیابی اطلاعات و پردازش زبان طبیعی در زبان فارسی مورد استفاده قرار می‌گیرند. مجموعه آزمون‌های قوانین در حوزه حقوق است و قوانین ایران را شامل می‌شود. یکی از پژوهش‌ها نیز از پرس‌وجوهای حوزه رایانه بهره برده است.

قضاوت ربط در پژوهش‌هایی که از مجموعه آزمون‌های استاندارد استفاده نموده‌اند با استفاده از قضاوت ربط انبوهه[2] انجام شده است و سه مورد از پژوهش‌ها که از نتایج بازیابی اسناد وب به‌عنوان مجموعه آزمون استفاده نمودند از قضاوت ربط انسانی کمک گرفته‌اند.

در همه این پژوهش‌ها بهبود عملکرد بازیابی گزارش شده است. اما از آنجایی که از مجموعه آزمون‌های مختلف و نیز از معیارهای متفاوت برای سنجش عملکرد ربط استفاده شده است، امکان مقایسه یافته‌ها با یکدیگر وجود ندارد.

1. WordNet
2. pooling

جدول ۳. منابع دانشی اصطلاحات گسترش و مجموعه آزمون و نوع قضاوت ربط

نویسنده (ها)	روش و الگوریتم گسترش پرس‌وجو	منبع دانشی اصطلاحات گسترش	مجموعه آزمون	قضاوت ربط
Saboori et al., (2008)	بازخورد ربط/ روکیو	مجموعه قوانین	قوانین	انبوهه
Dolamic & Savoy (2009)	بازخورد شبه ربط/ روکیو	مجموعه همشهری، مجموعه آزمون کِلِف	همشهری و کِلِف	انبوهه
Karisani et al., (2016)	بازخورد شبه ربط/ احتمالاتی	همشهری ۱ و ۲، مجموعه آزمون فایِر	همشهری ۱ و ۲، فایِر	انبوهه
Hashemi & Shakery (2014)	بازخورد شبه ربط/ احتمالاتی	مجموعه همشهری و اخبار بی‌بی‌سی	همشهری و اخبار بی‌بی‌سی	انبوهه
Aleahmad et al., (2007)	بازخورد شبه ربط/ تحلیل بافت محلی	مجموعه همشهری	همشهری	انبوهه
Hakimian & Taghiyareh, (2007)	بازخورد شبه ربط/ تحلیل بافت محلی	مجموعه همشهری	همشهری	انبوهه
Hakimian & Taghiyareh, (2008)	بازخورد شبه ربط/ تحلیل بافت محلی	مجموعه همشهری	همشهری	انبوهه
خالقی و مینایی (۱۳۹۴)	بازخورد شبه ربط/ تحلیل بافت محلی	مجموعه همشهری	همشهری	انبوهه
ساعدی (۱۳۹۰)	ساختارهای دانش/ مستقل از پیکره	فارسی‌نت	موتور کاوش	انسانی
دیانت و همکاران (۱۳۹۵)	ساختارهای دانش/ مستقل از پیکره	واژه‌نامه	فارس‌دات بزرگ	انبوهه
Farhoodi et al., (2009)	مبتنی بر وب/ پایگاه‌های آنلاین دانش	هستی‌شناسی مستخرج از ویکی‌پدیای فارسی	موتور کاوش گوگل	انسانی
خسروی و همکاران (۱۳۹۲)	مبتنی بر وب/ لاگ پرس‌وجو	موتور کاوش گوگل	موتور کاوش گوگل	انسانی
کریسانی (۱۳۹۰)	ترکیبی (ساختارهای دانش- مستقل از پیکره و بازخورد شبه ربط)	وردنت انگلیسی	همشهری	انبوهه
عبدالحسینی (۱۳۹۲)	ترکیبی (اسناد وب و ساختارهای دانش- مستقل از پیکره)	موتور کاوش گوگل و فارس‌نت	همشهری	انبوهه

بحث و نتیجه‌گیری

زبان فارسی یکی از زبان‌های مهم تکلم در خاورمیانه و آسیای میانه است و انتظار می‌رود تحقیقات در حوزه بازیابی اطلاعات در زبان فارسی توسعه یابد. یکی از حوزه‌هایی که به وظایف بازیابی اطلاعات کمک می‌کند گسترش پرس‌وجو است. در پژوهش حاضر به مرور نظام‌مند کارهای انجام شده برای گسترش پرس‌وجو در زبان فارسی با تمرکز بر روش‌ها، منابع دانشی و مجموعه آزمون‌ها پرداخته شده است.

در این مقاله روش‌های گسترش پرس‌وجو از دیدگاه منابع دانشی اصطلاحات گسترش دسته‌بندی شده است. زیرا منابع گسترش پرس‌وجو تضمین‌کننده اجرای روش‌ها و الگوریتم‌های گسترش پرس‌وجو هستند. مرور پژوهش‌ها نشان داد ۱۴ اثر به گسترش پرس‌وجوی زبان فارسی پرداخته‌اند. این پژوهش‌ها در قالب چهار روش تقسیم شدند: روش مبتنی بر ربط (هشت پژوهش)؛ مبتنی بر ساختارهای دانش (دو پژوهش)؛ مبتنی بر اطلاعات وب (دو پژوهش)؛ و مبتنی بر منابع ترکیبی (دو پژوهش). همه این پژوهش‌ها از نوع گسترش پرس‌وجوی خودکار هستند. دلیل اصلی محبوبیت روش‌های گسترش پرس‌وجوی خودکار این است که زمان و تلاش کمتری را از کاربران طلب می‌کنند.

در این مرور نشان داده شد که روش مبتنی بر ربط و به‌ویژه رویکرد بازخورد شبه ربط بیشتر از سایر روش‌ها برای گسترش پرس‌وجو در زبان فارسی مورد استفاده قرار گرفته‌اند همچنین سه تکنیک روکیو، مدل احتمالاتی و تحلیل بافت محلی در گسترش پرس‌وجوی زبان فارسی اعمال شده است. در همه تحقیقاتی که با استفاده از رویکرد بازخورد شبه ربط انجام شده است مجموعه آزمون استاندارد روزنامه همشهری به‌عنوان منبع دانشی اصطلاحات گسترش پرس‌وجو استفاده شده است. این در صورتی است که در تحقیقات انجام‌شده در زبان‌های دیگر مثل زبان انگلیسی اغلب از محتوای وب و جستجو در گوگل برای تأمین بافت پرس‌وجوها و انتخاب اصطلاحات گسترش استفاده می شود. این امر نشان می‌دهد نتایج وب و موتور کاوش گوگل در زبان فارسی هنوز مورد توجه و اعتماد پژوهشگران بازیابی اطلاعات قرار نگرفته است.

در روش مبتنی بر ساختارهای دانش دو اثر با رویکرد مستقل از پیکره و با استفاده از هستی‌شناسی فارس‌نت و یک واژه‌نامه فارسی انجام شده است. اما با رویکرد وابسته به پیکره، که اصطلاحات را از مجموعه‌هایی مانند خوشه‌بندی اصطلاحات و اصلاحنامه‌های خودکار مستخرج از متن انتخاب می‌کند پژوهشی در زبان فارسی انجام نشده است. استفاده از هستی‌شناسی‌ها جزو متأخرترین روش‌های گسترش

پرس‌وجو است و اغلب به‌منظور استنتاج بافت برای پرس‌وجوهای مبهم مورد استفاده قرار می‌گیرد. مفاهیم موجود در هستی‌شناسی‌ها را می‌توان برای رفع ابهام معنایی کلمه و نیز برای گسترش پرس‌وجو به‌کار برد. گسترش پرس‌وجوی خودکار با استفاده از دانش معنایی پیشرفت سریعی داشته است و خوش‌بینی زیادی در مورد ظرفیت‌های آن برای موفقیت در آینده وجود دارد. در پژوهش‌های زبان‌های مهم دنیا این رویکرد گسترش پرس‌وجوی خودکار بهبود چشمگیری در عملکرد بازیابی داشته است و تمایل به استفاده از این روش برای گسترش پرس‌وجو بیشتر از سایر روش‌ها است (Bhogal, et al., 2007). اما در زبان فارسی به‌دلیل نبود هستی‌شناسی مناسب استفاده از این روش مغفول مانده است.

در روش استفاده از اطلاعات مبتنی بر وب لازم است بیشتر به اسناد وب، پایگاه‌های آنلاین دانش (مانند ویکی‌پدیا) و لاگ پرس‌وجوها در زبان فارسی توجه شود. ویکی‌پدیا به‌عنوان یک منبع محبوب برای پژوهشگران در تحقیقات گسترش پرس‌وجوی زبان‌های دیگر مانند زبان انگلیسی به‌شمار می‌رود، چراکه بزرگ‌ترین دائره‌المعارف تحت وب است که مقالات آن به‌طور مرتب به‌روز می‌شوند و مقالات حوزه‌های جدید به آن افزوده می‌شود و با دارا بودن ویژگی‌های ساختاری و روابط معنایی منبع مفیدی برای پوشش نقاط ضعف ساختار و معنا است. در پژوهش‌های زبان فارسی توجه کافی به ویکی‌پدیا برای وظایف بازیابی اطلاعات و گسترش پرس‌وجو صورت نپذیرفته است و ابزار و پایگاهی مستخرج از ویکی‌پدیای فارسی برای امور مربوط به بازیابی اطلاعات تولید نشده است. استفاده از لاگ‌های پرس‌وجو به زبان فارسی نیز می‌تواند اطلاعات ارزشمندی از چگونگی رفتار کاربران در اصلاح و تغییر پرس‌وجوها را به‌منظور کمک به گسترش پرس‌وجوها نمایان سازد.

مرور پژوهش‌های گسترش پرس‌وجو در زبان فارسی نشان می‌دهد که اغلب آنها در حوزه مقالات خبری انجام شده‌اند. یکی از دلایل این امر این است که متون خبری با مقالات کوتاه و موضوع‌های ساده و مشخص نسبت به سایر حوزه‌ها، مورد اقبال پژوهشگران حوزه بازیابی اطلاعات هستند و در حوزه گسترش پرس‌وجو در زبان فارسی نیز این‌چنین است. همچنین از مجموعه آزمون همشهری در نیمی از پژوهش‌ها به‌عنوان منبع دانشی اصطلاحات گسترش و نیز مجموعه آزمون استفاده شده است. یکی دیگر از دلایل پرداختن به گسترش پرس‌وجو در حوزه مقالات خبری، می‌تواند وجود مجموعه آزمون استاندارد روزنامه همشهری در زبان فارسی باشد و نبود مجموعه آزمون‌های استاندارد در متون حوزه‌های دیگر مثل متون علمی، پژوهش‌ها را به‌سمت متون خبری سوق داده است.

مرور نظام‌مند پژوهش‌ها نشان می‌دهد گسترش پرس‌وجوها در زبان فارسی به‌صورت کلیدواژه‌ای انجام شده‌اند و استفاده از گسترش پرس‌وجو به‌صورت غیرکلیدواژه‌ای مانند گسترش عبارتی یا استفاده از اصطلاحات عمومی در وب مغفول مانده است.

در مجموع، مرور نظام‌مند پژوهش‌ها نشان داد، علاوه بر کمبود روش‌ها و تکنیک‌های به‌کار رفته در زبان فارسی، کمبود منابع دانشی اصطلاحات گسترش هم آشکار است. تعداد اندک آثار در گسترش پرس‌وجوی زبان فارسی با روش‌ها و رویکردهای محدود باعث شده که نتوان قضاوت کاملی درباره عملکرد و میزان بهبود بازیابی اطلاعات توسط آنها انجام داد. تحقیقات این حوزه نیازمند توسعه کمی است تا بتوان با تجزیه و تحلیل و ارزیابی روش‌های مختلف، درک دقیق‌تری از عملکرد آنها در زبان فارسی به‌دست آورده و از آن‌ها براساس نیاز در پایگاه‌های مختلف استفاده شود. همچنین به‌منظور مقایسه پژوهش‌های این حوزه لازم است از مجموعه آزمون استاندارد و معتبر با انواع معیارهای سنجش ربط برای بهبود عملکرد الگوریتم‌ها استفاده شود.

با توجه به شکاف‌های پژوهش‌های حوزه گسترش پرس‌وجو در زبان فارسی، پیشنهاداتی برای روش‌ها، منابع دانشی اصطلاحات گسترش و مجموعه آزمون‌ها جهت انجام پژوهش‌های آینده ارائه می‌شود.

بیش از نیمی از پژوهش‌های گسترش پرس‌وجو از روش مبتنی بر ربط استفاده کردند، تقریباً اکثر الگوریتم‌های مورد استفاده در روش مبتنی بر ربط در زبان فارسی پیاده‌سازی شده‌اند. اما به پیاده‌سازی روش‌های دیگر به این اندازه توجه نشده است و بهتر است پیاده سازی آنها نیز در زبان فارسی انجام شود، به‌ویژه، رویکردهای وابسته به پیکره در روش مبتنی بر ساختارهای دانش با استفاده از خوشه‌بندی اصطلاحات و اصلاحنامه‌های خودکار مستخرج از متن؛ استفاده از اسناد وب؛ استفاده از ویژگی‌های ساختاری و معنایی ویکی‌پدیا و پایگاه‌های مستخرج از آن؛ تحلیل لاگ‌های کاربران در نظام‌های بازیابی اطلاعات؛ موتورهای کاوش مختلف؛ و روش‌های ترکیبی که بیش از یک روش را برای گسترش پرس‌وجو در نظر می‌گیرند.

مرور پژوهش‌ها نشان می‌دهد استفاده از روابط ساختاری و معنایی هستی‌شناسی‌ها و ویکی‌پدیای فارسی چندان در پژوهش‌ها ظاهر نشده است. یکی از دلایل این امر نبود منابع مناسب برای تأمین اصلاحات گسترش پرس‌وجو است. فارس‌نت یک هستی‌شناسی عمومی است و تعداد واژگان آن پاسخ‌گوی انجام مطالعات بازیابی اطلاعات در حوزه‌های مختلف زبان فارسی نیست. لازم است برای تولید هستی‌شناسی‌های

معتبر در زبان فارسی برای امور بازیابی اطلاعات اقدام شود. همچنین تولید پایگاه‌های معتبر از ویکی‌پدیای فارسی می‌تواند منابع دانشی مفیدی را برای اعمال بازیابی اطلاعات به‌ویژه گسترش پرس‌وجو فراهم آورد.

بیش از نیمی از پژوهش‌های گسترش پرس‌وجو از روش مبتنی بر ربط استفاده نمودند و منبع تأمین اصطلاحات گسترش در این پژوهش‌ها مجموعه آزمون روزنامه همشهری و حوزه آنها متون خبری بوده است. لازم است مجموعه آزمون‌هایی استاندارد و معتبر در حوزه‌های دیگر به‌ویژه حوزه متون علمی تولید شود و منبع تأمین اصلاحات تحقیقات حوزه گسترش پرس‌وجو قرار گیرند.

منابع

خالقی، مرتضی، و مینایی، بهروز (۱۳۹۴). *چهارچوبی مستقل از زبان برای گسترش پرس‌وجو*. سومین کنفرانس بین‌المللی پژوهش‌های کاربردی در مهندسی کامپیوتر و فن‌آوری اطلاعات.

خسروی، عبدالرسول، فتاحی، رحمت‌الله، پریرخ، مهری و دیانی، محمدحسین (۱۳۹۲). بررسی کارآمدی کلیدواژه‌ها و عبارت‌های پیشنهادی موتور کاوش گوگل در بسط جستجو و افزایش ربط از دیدگاه دانشجویان تحصیلات تکمیلی. *پژوهش‌های نظری و کاربردی در علم اطلاعات و دانش‌شناسی*، ۳(۱)، ۱۳۳-۱۵۰.

دیانت، روح‌الله، علی‌احمدی، مرتضی، اخلاقی، محمدیحیی و باباعلی، باقر (۱۳۹۵). ارائه یک روش جدید بازیابی اطلاعات مناسب برای متون حاصل از بازشناسی گفتار. *پردازش علائم و داده‌ها*، ۴(۳). ۹۳-۱۰۸.

ساعدی، سیامک (۱۳۹۰). *گسترش پرس‌وجو در موتورهای جستجوی فارسی*. پایان‌نامه کارشناسی ارشد رشته مهندسی فناوری اطلاعات- شبکه‌های کامپیوتری. دانشگاه یزد، دانشکده مهندسی برق و کامپیوتر.

شبان‌زاده حبیب‌آبادی، مژگان (۱۳۸۹). *گسترش معنایی پرس‌وجو*. پایان‌نامه کارشناسی ارشد دانشگاه اصفهان، دانشکده فنی و مهندسی.

عبدالحسینی، زهرا (۱۳۹۲). *بسط پرس‌وجوی کاربر با بهره‌گیری از روش‌های استنتاج روابط معنایی در بانک‌های اطلاعاتی متنی*. پایان‌نامه کارشناسی ارشد، دانشکده فنی، دانشگاه الزهرا.

کریسانی، پیام (۱۳۹۰). *گسترش پرس‌وجوهای فارسی در موتورهای جستجو*. پایان‌نامه کارشناسی ارشد رشته کامپیوتر- گرایش نرم‌افزار. دانشگاه تهران، دانشکده مهندسی برق و کامپیوتر.

ملبوس‌باف، رامین و عزیزی، فریدون (۱۳۸۹). مرور سیستماتیک "Systematic Review" چیست و چگونه نگاشته می‌شود؟. *پژوهش در پزشکی*، ۳۴(۳)، ۲۰۳-۲۰۷.

Abdelmgeid Amin, A. (2008). Using a query expansion technique to improve document retrieval. *Information Technologies and Knowledge*, 7(2), 343-345.

Agichtein, E., & Cucerzan, S. (2005). *Predicting Extraction Performance Using Context Language Models*. In the SIGIR 2005 Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-World Applications, 2005.

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., & Oroumchian, F. (2009). Hamshahri: A standard Persian text collection. *Knowledge-Based Systems*, 22(5), 382-387.

AleAhmad, A., Hakimian, P., Mahdikhani, F., & Oroumchian, F. (2007, February). N-gram and local context analysis for persian text retrieval. In *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on* (pp. 1-4). IEEE.

Atwan, J., & Mohd, M. (2017). Arabic Query Expansion: A Review. *Asian Journal of Information Technology*, 16(10), 754-770.

Azad, H. K., & Deepak, A. (2019). Query expansion techniques for information retrieval: a survey. *Information Processing & Management*, *56*(5), 1698-1735.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison-Wesley Harlow, England.

Bendersky, M., & Croft, W. B. (2008, July). Discovering key concepts in verbose queries. *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 491-498). ACM.

Bhogal, J., MacFarlane, A., & Smith, P. (2007). A review of ontology based query expansion. *Information processing & management*, 43(4), 866-886.

Dolamic, L., & Savoy, J. (2009, September). Ad hoc retrieval with the Persian language. *In Workshop of the Cross-Language Evaluation Forum for European Languages* (pp. 102-109). Springer, Berlin, Heidelberg.

Efthimiadis, E. N. (1996). Query Expansion. *Annual review of information science and technology (ARIST)*, 31, 121-87.

Farhoodi, M., Mahmoudi, M., Bidoki, A. Z., Yari, A., & Azadnia, M. (2009). Query expansion using persian ontology derived from Wikipedia. *World Applied Sciences Journal*, *7*(4), 410-417.

Hakimian, P., & Taghiyareh, F. (2007, December). Tuning Local Context Analysis for Farsi Documents. In *Semantic Media Adaptation and Personalization, Second International Workshop on* (pp. 116-121). IEEE.

Hakimian, P., & Taghiyareh, F. (2008, December). Customizing local context analysis for farsi information retrieval by using a new concept weighting algorithm. *In 2008 Third International Workshop on Semantic Media Adaptation and Personalization* (pp. 45-51). IEEE.

Harman, D., & Buckley, C. (2004, July). The NRRC reliable information access (RIA) workshop. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 528-529). ACM.

Hashemi, H. B., & Shakery, A. (2014). Mining a Persian–English comparable corpus for cross-language information retrieval. *Information Processing & Management*, 50(2), 384-398.

Karisani, P., Rahgozar, M., & Oroumchian, F. (2016). A query term re-weighting approach using document similarity. *Information Processing & Management*, *52*(3), 478-489.

Lavrenko, V., & Croft, W. B. (2017, August). Relevance-based language models. *In ACM SIGIR Forum* (Vol. 51, No. 2, pp. 260-267). ACM.

Lee, K. S., Croft, W. B., & Allan, J. (2008, July). A cluster-based resampling method for pseudo-relevance feedback. *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 235-242). ACM.

Mehdi, M., Okoli, C., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2017). Excavating the mother lode of human-generated text: A systematic review of research that uses the wikipedia corpus. *Information Processing & Management*, 53(2), 505-529.

Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research. *Sprouts*, 10-26.

Phan, N., Bailey, P., & Wilkinson, R. (2007, July). Understanding the relationship of information need specificity to search query length. *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 709-710). ACM.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3), 129-146.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing*, 313-323.

Saboori, F., Bashiri, H., & Oroumchian, F. (2012). Assessment of query reweighing, by rocchio method in farsi information retrieval. *International Journal of Information Science and Management (IJISM)*, *6*(1), 9-16.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., ... & Assi, S. M. (2010). Semi automatic development of farsnet; the persian wordnet. *In Proceedings of 5th global WordNet conference, Mumbai, India* (Vol. 29).

Spink, A., Wolfram, D., Jansen, M. B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.

Wollersheim, D. (2005). *Dynamic query expansion for information retrieval of imprecise medical queries*. La Trobe University.

Zhang, H. (2013). *Query enhancement with topic detection and disambiguation for robust retrieval*. Indiana University.